UN ANÁLISIS COMPARATIVO DE UNA SVM Y UN MODELO LOGIT EN UN PROBLEMA DE CLASIFICACIÓN DE ASEGURADOS

Antonio Heras Martínez[1], Piedad Tolmos Rodríguez-Piñero[2], Julio Hernández-March[2]

[1] Departamento de Economía Financiera y Contabilidad I. Facultad de CC Económicas y Empresariales. Universidad Complutense de Madrid.
[2] Departamento de Economía Financiera y Contabilidad II. Facultad de Ciencias Jurídicas y Sociales. Universidad Rey Juan Carlos.
emails: antonio.heras@ccee.ucm.es, piedad.tolmos@urjc.es, julio.hernandez.march@urjc.es

Palabras Clave.- Clasificación de Asegurados del Seguro del Automóvil, Factores de Riesgo, Máquinas de Vectores Soporte, Algoritmos Genéticos, Modelo Logit.

Resumen.- Con este artículo se pretende realizar una aproximación a la clasificación de los asegurados de una cartera de una compañía del seguro del automóvil atendiendo a si han presentado o no siniestro en un año[1]. Para realizar la clasificación utilizaremos una técnica de Aprendizaje conocida como Máquina de Vectores Soporte. En aras de preservar la capacidad de generalización del clasificador, realizaremos además una selección de los factores de riesgo que describen a los asegurados de la cartera, escogiendo los más relevantes de cara a la siniestralidad. Para ello emplearemos de nuevo herramientas de Aprendizaje Máquina, esta vez Algoritmos Genéticos. Se ejecutarán varios experimentos, comparando la tasa de clasificación obtenida utilizando todos los factores de riesgo, y sólo los seleccionados. También se compararán los mejores resultados conseguidos con la clasificación lograda por el modelo *logit*, que nos permitirá analizar hasta qué punto son comparables las técnicas del Aprendizaje Máquina y los modelos estadísticos utilizados habitualmente en la resolución de este tipo de

[1] Los datos empleados en los experimentos han sido cedidos por la aseguradora MAPFRE en el marco de la Beca de Riesgos y Seguros 2006 que nos concedió su Fundación MAPFRE Estudios.
Este artículo ha sido recibido en versión revisada el 13 de julio de 2010.

problemas. Aprovecharemos, además, la salida del *logit* para comparar los factores de riesgo que resultan más relevantes con los que se seleccionaron a través del Algoritmo Genético (AG). Los resultados obtenidos son alentadores, probando que las técnicas de aprendizaje, y las SVM en particular, pueden resultar muy útiles para resolver problemas de clasificación en seguros.

Key words.- Insurance Automobile Policies Classification, Risk Factors, Support Vector Machines, Genetic Algorithms, Logit Model.

Abstract.- In this paper we propose a new approach for classifying the policies of an insurance automobile company according to the prediction of their claims for the next year. We use for this purpose sets of risk factors obtained from one database of an important Spanish insurance company[2]. Our approach is based on the Learning Machines methodology. The algorithm we suggest is based on the application of a standard Support Vector Machine (SVM), hybridized with a Genetic Algorithm. The SVM is used to classify the policies as failed (reporting claims) or not failed, according to their risk factors, whereas the GA is used to perform a pre-selection in the risk factors space of the SVM. We will do several experiments, comparing the obtained classification rate including all the risk factors, with that including just the selected risk factors. We'll also compare this results with those obtained using the Logit model (both classification rate and selected risk factors), allowing us to analyze if this Learning Machines are comparable to statistical techniques commonly used to solve this kind of problems. The obtained results are very encouraging and show that learning techniques in general and SVM in particular, can be useful tools for solving classification problems in insurance.

1. Introducción.

Cuando el número de pólizas en una compañía aseguradora es lo suficientemente grande, el desarrollo de un sistema de clasificación adecuado es el primer paso para lograr una prima justa. Se trata de clasificar a sus clientes del modo más homogéneo posible atendiendo al riesgo, de

[2] Data used in the experiments were given up by MAPFRE Insurance Company, within the Risk and Insurance Grant 2006 of MAPFRE Studies Foundation.

manera que los asegurados pertenecientes a un mismo grupo paguen idéntica prima.

En la literatura, se pueden encontrar soluciones a esta tarea empleando métodos estadísticos. Las técnicas del análisis estadístico multivariante son las que permiten organizar procesos de selección, teniendo en cuenta simultáneamente el conjunto de factores de riesgo.

Técnicas tales como las Redes Neuronales Artificiales, o las Máquinas de Vectores Soporte han demostrado ser unos clasificadores excelentes en términos de la llamada *tasa de clasificación* y se han aplicado con éxito en multitud de problemas complejos, como el del reconocimiento de caracteres escritos, la "limpieza" de imágenes, minería de datos, diagnósticos médicos, etc.

Vamos por tanto a resolver el problema valiéndonos de herramientas tomadas de lo que se conoce como Aprendizaje Máquina. El planteamiento concreto que haremos será el siguiente: dada una cartera de asegurados de una conocida empresa del seguro del automóvil, descritos por sus factores de riesgo, pretendemos clasificarlos en dos clases, atendiendo a si han presentado o no siniestros en el periodo de un año. Como luego veremos, un problema de clasificación de este estilo lleva aparejado un segundo problema, el de la Selección de Características (factores de riesgo). Efectivamente, con el objeto de mejorar la capacidad de generalización del clasificador, su estabilidad, y el tiempo de computación, a menudo es necesario, si el nº de variables es grande, realizar una selección de las variables importantes, las que retienen la mayor cantidad de información. Esta cuestión tiene en nuestro caso un valor añadido, en cuanto al interés que la información sobre los factores de riesgo realmente relevantes de cara a la siniestralidad, pueda tener para la aseguradora.

Así, comenzaremos por formular matemáticamente el problema, estableciendo a continuación qué se entiende por Aprendizaje y en qué difieren básicamente éstos métodos de las técnicas estadísticas clásicas. Describiremos brevemente las técnicas que hemos empleado para solucionar los problemas y pasaremos entonces a la ejecución de los experimentos, comentando los resultados alcanzados. Aplicaremos una Máquina de Vectores Soporte[3] para clasificar y un Algoritmo Genético para seleccionar los factores. Compararemos los resultados de la clasificación con los obtenidos aplicando una técnica estadística válida para abordar este tipo de

[3] En adelante se utilizará la abreviatura correspondiente a las siglas en inglés, SVM.

cuestiones, el modelo logit, extendiendo así los resultados alcanzados en la Tesis doctoral de Piedad Tolmos en la que se empleó la técnica del Análisis Discriminante, menos potente que el actual logit. Asimismo, contrastaremos la Tasa de Clasificación alcanzada por la SVM utilizando todos los factores de riesgo y sólo los seleccionados por el AG. Emplearemos además la salida del logit para comparar los factores de riesgo que resultan más relevantes con los que se seleccionaron a través del Algoritmo Genético (AG).

Finalizaremos con las conclusiones que se extraen de lo presentado en el artículo, y comentando algunas de las aplicaciones que estamos realizando en la actualidad.

2. Formalización del problema.

El tipo de problema que planteamos se puede englobar dentro de lo que se conoce como *problemas de clasificación con múltiples atributos* cuya tarea básica es asignar un *objeto*, descrito por los valores que toman ciertos *atributos*, a una serie de *clases*.

Matemáticamente, la representación de un problema de clasificación con múltiples atributos es la siguiente [Schölkopf, 1999]: se quiere estimar una función (de decisión) $f : \Re^n \to \{\pm 1\}$ empleando datos (observaciones u *objetos*) del conjunto de observaciones que se utilizarán para *entrenar* lo que se conoce como "máquina de clasificación" (red neuronal, algoritmo genético,…). Consideraremos para ello una serie de objetos $\{\mathbf{x}_i\}, \mathbf{x}_i \in \Re^n$, $i \in \{1,...l\}$ generados por cierta función de distribución de probabilidad desconocida $P(x,y)$, donde $y_i \in \{1,-1\}$ constituyen el conjunto de "etiquetas" asociadas (es la salida, la que indica a qué clase pertenece cada $\mathbf{x}_i$). De este modo, el conjunto de datos considerados sería

$$(\mathbf{x}_1, y_1),...,(\mathbf{x}_l, y_l) \in \Re^n \times \{\pm 1\}$$

y el objetivo es que la función f clasifique correctamente los ejemplos nuevos que se la presenten ($\mathbf{x}$,y), esto es, que $f(\mathbf{x})$=y para ejemplos ($\mathbf{x}$,y) generados por la misma distribución de probabilidad *"subyacente"* $P(\mathbf{x},y)$ que los datos utilizados para el entrenamiento.

El *error de entrenamiento* se puede definir como:

$$R_{ent}[\alpha] = \frac{1}{l}\sum_{i=1}^{l}\frac{1}{2}\left| f(\mathbf{x_i}, \alpha) - y_i \right| \quad (1)$$

El error de test (*riesgo*) esperado para una máquina de entrenamiento se define como

$$R[\alpha] = \int \frac{1}{2}\left| f(\mathbf{x}, \alpha) - y \right| dP(\mathbf{x}, y) \quad (2)$$

La cantidad $\frac{1}{2}\left| f(\mathbf{x}, \alpha) - y_i \right|$ recibe el nombre de *pérdida*.

2.1 El problema de la clasificación de los asegurados

Para predecir la siniestralidad de un asegurado vamos a separar a todos los clientes en dos clases: la de los que tendrán siniestros, y la de los que no. Por ello, lo trataremos como un problema de clasificación con múltiples atributos de tipo "simple", esto es, con sólo dos clases. Los conjuntos $\{\mathbf{x}_i\}, i \in \{1,...l\}\ \mathbf{x}_i \in \Re^n$ representan a los asegurados, descritos por un conjunto de *n* factores de riesgo (cada componente de $\mathbf{x}_i$, $\mathbf{x}_{ij}$, es un factor), y las etiquetas $y_i \in \{-1,1\}$ indicarían la clase, -1 si no presentan siniestros, y 1 en caso contrario. De este modo, durante el periodo de entrenamiento estaríamos manejando pares del tipo $(\mathbf{x_i}, y_i)$, con y_i conocida; el objetivo será, recordemos, que se clasifiquen correctamente los ejemplos nuevos que se presenten $(\mathbf{x}, y)$ generados por la misma distribución de probabilidad *"subyacente"* $P(\mathbf{x}, y)$ que los datos utilizados para el entrenamiento del clasificador.

En el caso de la selección de factores, la entrada para el clasificador será la misma, el conjunto $\{\mathbf{x}_i\}, i \in \{1,...l\}\ \mathbf{x}_i \in \Re^n$, pero la salida será $\{\mathbf{x}_i\}, i \in \{1,...l\}\quad \mathbf{x}_i \in \Re^m\ m < n$, con un error de clasificación menor.

2.2 Los datos utilizados

Para la ejecución hemos empleado una muestra de 58237 asegurados, obtenida al enlazar la cartera de Clientes con la de Siniestros del año 2003 que nos proporcionó MAPFRE.

Los factores de riesgo que describían a cada asegurado eran originalmente 11: Antigüedad del carnet, Edad, Tipo de carnet, Sexo, Estado Civil, Profesión, Antigüedad del vehículo, Uso, Zona de Circulación, Potencia, Valor.

Sin embargo, hubo que segregarlos para obtener variables categóricas, como exigía el sistema, manejando finalmente 105 variables de entrada. La salida, recuérdese, sólo tomaba dos valores, 1 (siniestro) o -1 (no siniestro).

3. Las técnicas empleadas

3.1 El Aprendizaje Máquina

El enfoque estadístico "clásico" para abordar un problema de clasificación como el que nos planteamos, en el que se cuenta con una gran cantidad de datos, consiste en asumir que tales datos están generados por una *distribución de probabilidad* subyacente que nos es desconocida y a partir de la cual diseñaremos el *clasificador*. Sin embargo, existen otras aproximaciones al problema, como la que nos proponen Vapnik y otros autores de la Teoría del Aprendizaje. La idea básica es diseñar el clasificador directamente desde los datos mediante determinados algoritmos, que en su caso se basan en esta Teoría del Aprendizaje.

Este modo de plantear el problema nos conducirá a la necesidad de analizar la información que comprenden los grandes conjuntos de datos. La habilidad de extraer el conocimiento que se encuentra escondido entre esos datos, y utilizarlo convenientemente, está teniendo una importancia creciente en el mundo contemporáneo. Las aproximaciones más recientes para desarrollar modelos a partir de los datos se han inspirado en las capacidades de *aprendizaje* de los sistemas biológicos y, en particular, en las de los humanos. De hecho, los sistemas biológicos aprenden a hacer frente a la desconocida naturaleza estadística del entorno conducidos por los datos. Los humanos, como los animales, tienen la capacidad superior de *reconocer patrones*, como las de identificar caras, voces u olores. El campo del reconocimiento de patrones tiene como objetivo el construir sistemas

artificiales que imiten las habilidades de reconocimiento de los humanos, y avanza hacia él basándose en principios de ingeniería y estadística.

En un sentido amplio, cualquier método que incorpore información de las muestras de entrenamiento en el diseño de un clasificador emplea *aprendizaje*. La creación de clasificadores implica el planteamiento de una forma general de modelo, o de clasificador, y el uso de los datos de entrenamiento para *aprender* o estimar los parámetros desconocidos del modelo. Cuando hablemos aquí de *aprendizaje* lo haremos como una forma de algoritmo para reducir el error sobre el conjunto de entrenamiento. Concretamente, un *algoritmo de aprendizaje* es aquel que toma los datos de entrenamiento como entrada (*input*) y selecciona la hipótesis (la función candidata de entre todas a ser la que relaciona las salidas con las entradas, esto es, la *función de decisión*) de entre todas las posibles.

3.2 Las Máquinas de Vectores Soporte

La SVM es una técnica de clasificación que ha demostrado sobradamente su capacidad de resolución frente a problemas de elevado grado de complejidad. Diseñada en principio para tratar problemas de clasificación binarios (en dos grupos), se trata de una máquina de aprendizaje que implementa la siguiente idea: cuando no sea posible separar los datos en el espacio de entrada con un hiperplano lineal, trasladar, mediante una aplicación no lineal, los vectores de entrada a un nuevo espacio de dimensión más alta. En este nuevo espacio se construirá una superficie de decisión lineal. Las especiales propiedades que poseerá esta superficie garantizarán que la capacidad de generalización de la máquina de aprendizaje sea alta. Aunque esta idea se empleó en los primeros experimentos para datos que podían separarse sin errores, se puede extender para el caso no separable con notable éxito. La parte conceptual del problema la resolvió Vapnik para el caso de *hiperplanos óptimos* para clases separables. En este contexto, Vapnik definió un hiperplano óptimo como una función de decisión lineal con el margen de separación máximo entre los vectores de las dos clases. Se observó entonces que para construir tal hiperplano, uno sólo debía tener en cuenta una cantidad pequeña de los datos de entrenamiento, los llamados *vectores soporte*, quienes determinaban ese *margen*.

Sea un conjunto de asegurados representado por sus factores de riesgo expresados mediante los vectores $\{\mathbf{x}_i\}, i \in \{1,...l\}$, y un conjunto de etiquetas

asociadas $y_i \in \{-1,1\}$[4] que determinan a qué clase pertenece cada asegurado. Supongamos que es posible separar este conjunto de entrenamiento mediante un hiperplano lineal. Los puntos $\mathbf{x}$ que pertenecen al hiperplano satisfacen la ecuación $\mathbf{w} \cdot \mathbf{x} + b = 0$, donde $\mathbf{w}$ es un vector normal al hiperplano, y $|b|/\|\mathbf{w}\|$ es la distancia perpendicular del hiperplano al origen (tomamos $\|\cdot\|$ como la norma euclídea). De entre todos los hiperplanos capaces de separar los datos, existe un único *hiperplano óptimo*, en el sentido de que es capaz de separar los puntos con el mayor margen de separación entre cada elemento del conjunto de entrenamiento y el hiperplano. En este sentido, el algoritmo de aprendizaje diseñado por Vapnik y Chervonenkis, la SVM, resuelve el siguiente problema:

"Encontrar $\mathbf{w} \in \Re^n$ y $b \in \Re$ que minimicen $\tau(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$

sujeto a $y_i\left(\mathbf{w}^t \cdot \mathbf{x}_i + b\right) \geq 1 \qquad \forall\, i = 1,...,l$ " (3)

Cuando se hallen $\mathbf{w}$ y *b*, la regla de clasificación para los asegurados será, simplemente, $sign\left(\mathbf{w}^t \cdot \mathbf{x}_i + b\right)$[5], y el error de clasificación cometido vendrá dado por $R_{ent}(\mathbf{w}, b)$, tal y como se describió anteriormente. Por otro lado, aquellos puntos que verifican la igualdad en la inecuación $y_i\left(\mathbf{w}^t \cdot \mathbf{x}_i + b\right) \geq 1$, y cuya eliminación cambiaría la solución que encontremos, son los que llamaremos *vectores soporte* . Estos vectores, pertenecerán a uno de los dos posibles hiperplanos óptimos de separación de los que hablábamos antes, representados por las ecuaciones $\mathbf{w}^t \cdot \mathbf{x}_i + b = 1$ para un hiperplano, y $\mathbf{w}^t \cdot \mathbf{x}_i + b = -1$ para el otro.

Si tratamos de aplicar el algoritmo anterior a datos no separables, no encontraremos ninguna solución factible, pues la función objetivo crece desmesuradamente. Para evitarlo, se relaja la restricción 2 cuando sea necesario, lo que se logra mediante la introducción de unas variables nuevas de pequeño tamaño. La formulación del problema queda ahora:

---

[4] Describimos el caso más sencillo de dos únicas clases, pues para el general de *K* clases basta con tomar, como ya se ha visto y_{ik} en vez d y_i.

[5] Obsérvese que se corresponde con las funciones de decisión $f(\mathbf{x}) = sign\left(\mathbf{w}^t \cdot \mathbf{x}_i + b\right)$ que describimos en el punto 2.

" Encontrar $\mathbf{w} \in \Re^n$ y $b \in \Re$ y ξ_i $i = 1,\ldots,l$ que

$$\text{Minimicen} \quad \tau(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\xi_i \tag{4}$$

$$\text{Sujeto a} \quad y_i\left(\mathbf{w}^t \cdot \mathbf{x}_i + b\right) \geq 1 - \xi_i \quad \text{y} \quad \xi_i > 0 \quad \forall\, i = 1,...,l \text{ "} \tag{5}$$

donde *C* es un parámetro que el clasificador deberá estimar.

Por último, en el caso de la SVM no lineal, se proyectan las variables de entrada en un espacio de dimensión mayor (normalmente de dimensión infinita) que aquel al que pertenecían dichas variables, y se aplica la SVM descrita anteriormente en este nuevo espacio, conocido como *espacio de características*. De este modo, la SVM no lineal es capaz de separar los asegurados con una probabilidad de error dada por $R_{ent}(\mathbf{w}, b)$. Esa proyección se realiza utilizando las funciones *núcleo* (kernel).

3.3 La selección de factores

El llamado "problema de selección de características", esto es, la selección de los factores o rasgos que permitan desechar aquellos elementos que se revelen como irrelevantes para el estudio que se desea realizar, ha resultado ser de especial importancia en la mayoría de los problemas de aprendizaje supervisado.

En los problemas de clasificación como el que nos ocupa, el objetivo es seleccionar un subconjunto de variables de entrada (factores de riesgo) que sean los que preserven o mejoren la capacidad del clasificador [Weston et al. (2000)].

Entre los distintos modos de tratar este problema, el más frecuente es el siguiente: dado un conjunto de datos $(\mathbf{x_1}, y_1),...,(\mathbf{x}_l, y_l)$, con $\mathbf{x}_i \in \Re^n$ y $y_i \in \{-1,1\}$, extraer un subconjunto de *m* variables (*m* < *n*) que posean el error de clasificación menor [ibídem].

En nuestro caso, seguiremos a Weston et al. para seleccionar los mejores factores de riesgo (componentes de los vectores $\{\mathbf{x}_i\}, i \in \{1,...l\}$), esto es, aquellos que realmente describan el estado del asegurado, y eliminaremos

aquellos factores redundantes o irrelevantes. De este modo mejoraremos el funcionamiento de la SVM, que trabajará en el proceso de clasificación sólo con los factores de riesgo que hayamos reservado.

Existen varias técnicas para resolver el problema de la selección de características. En este caso hemos escogido los Algoritmos Genéticos.

3.4 Algoritmos Genéticos

Los algoritmos genéticos son un logro más de la Inteligencia Artificial en su intento de replicar comportamientos biológicos mediante la computación. Se trata de algoritmos de búsqueda basados en la mecánica de la selección natural y de la genética. Utilizan la información histórica para encontrar nuevos puntos de búsqueda de una solución.
Se puede pensar en cada "*cromosoma*" de un algoritmo genético como en un punto en el espacio de búsqueda de candidatos a soluciones. El algoritmo genético procesa poblaciones de cromosomas, reemplazando sucesivamente cada población por otra. El algoritmo suele requerir una *función de capacidad* o potencial que asigna una puntuación (la capacidad) a cada cromosoma de la población actual. La capacidad o el potencial de un cromosoma depende de cómo resuelva ese cromosoma el problema a tratar.

La forma más simple de algoritmo genético utiliza tres tipos de operadores: selección, cruce y mutación.

Selección o reproducción: Este operador escoge cromosomas entre la población para efectuar la reproducción. Cuanto más capaz sea el cromosoma, más veces será seleccionado para reproducirse.

Cruce: Se trata de un operador cuya labor es elegir un lugar, y cambiar las secuencias antes y después de esa posición entre dos cromosomas, para crear nueva descendencia (por ejemplo, las cadenas 10010011 y 11111010 pueden cruzarse después del tercer lugar para producir la descendencia 10011010 y 11110011). Imita la recombinación biológica entre dos organismos haploides.

Mutación: Este operador produce variaciones de modo aleatorio en un cromosoma (por ejemplo, la cadena 00011100 puede mutar su segunda posición para dar lugar a la cadena 01011100). La mutación puede darse en cada posición de un bit en una cadena, con una probabilidad, normalmente muy pequeña (por ejemplo 0.001).

Cada iteración del algoritmo recibe el nombre de generación. Lo usual es iterar el proceso de 50 a 500 o más veces. El conjunto completo de generaciones se llama serie. Al concluir una serie, a menudo se encuentran entre la población uno o más cromosomas con elevada capacidad.

En nuestro problema de selección de los factores relevantes para la tarificación a priori, la población del AG está integrada por un número ξ de cadenas binarias $\sigma \in \{0,1\}^n$ a las que se aplica el procedimiento iterativo de los operadores genéticos. Una componente $\sigma_i = 1$ equivale a afirmar que el factor de riesgo correspondiente debe ser tenido en cuenta para la SVM, mientras que si $\sigma_i = 0$, se eliminará ese factor del conjunto de factores. Debe observarse que cada individuo de la población del AG (un vector σ) permanece para un conjunto de factores diferente de la SVM. La función de capacidad asociada a cada individuo es el error de clasificación obtenido al clasificar los puntos de entrenamiento $(\mathbf{x} * \sigma, y)$, que será estimado por $R_{ent}(\mathbf{w}, b, \sigma)$. Una última apreciación: como el AG maximiza la función de capacidad, y el objetivo en un problema de selección de características es minimizar la probabilidad de error, se introducirá una función de capacidad modificada,

$$F = 100(1- R_{ent}(\mathbf{w}, b, \sigma)) \tag{6}$$

3.5 El modelo logit

Considérese, por otro lado, que se quiera explicar la ocurrencia aleatoria del siniestro como consecuencia de un conjunto de características x_j para $j = \{1,...,k\}$ relativas al conductor y de un elemento debido al azar:

$$Y_i = \sum_{j=1}^{k} \beta_j x_{ji} + u_i \tag{7}$$

La naturaleza dicotómica de la variable dependiente Y (1 si el conductor asegurado ha sufrido un siniestro y -1 en caso contrario) obliga al empleo de un modelo no lineal que la relacione con las variables explicativas x_j (las mismas que se han empleado en el modelo SVM). Asumiendo que la

perturbación aleatoria u sigue una distribución logística con media nula y varianza constante, el modelo quedaría (Aldrich & Nelson, 1986):

$$E(Y_i) = P(Y_i = 1) = \frac{e^{Z_i}}{1 + e^{Z_i}} \tag{8}$$

donde:

$$Z_i = \sum_{j=1}^{k} \beta_j x_{ji} = \ln(\frac{P_i}{1 - P_i}) \equiv \log it \tag{9}$$

Expresión en la que el cociente de probabilidades se conoce como "odds". El cociente entre dos odds se conoce como ratio de odds y permite medir el efecto multiplicativo que tiene un aumento unitario de cualquiera de las variables explicativas x_j sobre la "odds" de tener un siniestro (Liao, 1994). En nuestro caso, dicho ratio sirve para medir el riesgo de siniestralidad al cambiar el valor de una variable, cuando el resto de las variables permanecen fijas:

$$x_{ij}^{'} = x_{ij} + 1 \Rightarrow \frac{\left[\frac{P_{id}}{1 - P_i}\right]_{x_{ij}^{'}}}{\left[\frac{P_i}{1 - P_i}\right]_{x_{ij}}} = e^{\beta_j} \tag{10}$$

El efecto de una variable sobre la probabilidad de ocurrencia de un siniestro también vendrá informado por el estadístico de Wald W_j, que permite criticar la validez de la estimación puntual del parámetro β_j que pondera a la variable, en función de su dispersión (Hernández-March, 2003):

$$W_j = \left(\frac{\widehat{\beta_j}}{S_{\widehat{\beta_j}}} \right)^2 \tag{11}$$

Cuanto mayor sea W_j más precisa será la estimación de β_j, de ahí que ante dos variables con coeficientes beta significativos (*p-valor* menor que .1), se preferirá aquel que tenga mayor estadístico de Wald.

4. Resolución del problema

4.1Clasificación de los asegurados utilizando todos los factores de riesgo

En este experimento utilizamos un software gratuito para la SVM llamado Libsvm, con un *kernel* de base radial, $k(\mathbf{x},\mathbf{x}_i) = \exp\left(-\left\|\mathbf{x}-\mathbf{x}_i\right\|^2 / c\right)$. Escogimos como método para determinar el error en el conjunto de Test el de la *Validación Cruzada*[6] en 5 *pliegues.*

Los resultados se presentan en forma de la Tasa de Clasificación [(Nº de aciertos) / (Nº de casos)] y de la Matriz de Confusión, en cuya diagonal principal aparecen los casos acertados, y en la secundaria los errores cometidos en cada clase.

Comparación del modelo logit con la SVM

En lo que se refiere a la capacidad predictiva del modelo, la tabla 1 recoge las tasas de clasificación que se han obtenido al aplicar los dos modelos objeto de comparación: una SVM y el logit[7].

Tabla 1: Tasas de clasificación obtenidas en los dos modelos empleados

	SVM			LOGIT		
Tasa de clasificación	77.72%.			70.8%		
Matriz de Confusión (en porcentajes)		-1	1		-1	1
	-1	76.41%	23.59%	-1	71.7%	28.3%
	1	20.87%	79.13%	1	30.1%	69.9%

[6] Dividir el conjunto de datos de entrada en n subconjuntos, entrenar el modelo con n-1 de los n conjuntos, y validar los resultados con el conjunto restantes. Repetir el proceso para cada uno de las n posibles elecciones del conjunto omitido.

[7] Para pronosticar los siniestros en el modelo logit se tomó como valor de corte 0,514 (que es la proporción muestral de asegurados que declaró siniestro).

En este sentido, se puede observar que el modelo SVM pronostica correctamente el 77.72% de los casos observados, casi 7 puntos por encima de lo que lo hace el modelo logit (diferencia que se reduce a algo menos de 5 puntos en el caso de los asegurados que no han tenido siniestro, pero que se incrementa por encima de los 9 puntos en el apartado de los que sí han sufrido accidente).
Por otra parte, los porcentajes que se logran con el modelo logit mejoran los alcanzados con el Análisis Discriminante (Bousoño, Heras y Tolmos 2008) en más de 2 puntos en el caso (-1,-1), si bien empeoran en el (1,1) – 71.8% del AD frente al 69.9% del logit –.

4.2 Selección de factores de riesgo

Resultados obtenidos por el modelo logit

La tabla 2 recoge los resultados de la estimación por máxima verosimilitud de los parámetros del modelo, empleando el programa SPSS. En lo que se refiere a la bondad del ajuste puede observarse que el estadístico Chi-cuadrado es muy significativo, lo que permite rechazar la hipótesis nula de que todos los parámetros del modelo, excepto el término independiente, sean nulos. El estadístico R cuadrado de Nagelkerke hay que interpretarlo en el mismo sentido.

En lo que respecta a la estimación de los parámetros, se observa que la región en la que el asegurado conduce habitualmente es la variable que mejor discrimina a la hora de explicar la siniestralidad. Dentro de este apartado, los conductores que corren más riesgo son los de Madrid y Barcelona. En concreto, el riesgo de declarar un accidente en Madrid es 7.803 veces mayor al de hacerlo en Andalucía sin incluir Sevilla, que es la categoría de referencia, mientras que en Barcelona ese riesgo es 7.45 veces mayor. Otras zonas de conducción en las que resulta más probable declarar un siniestro que en Andalucía sin Sevilla son: Castilla y León, Galicia, Valencia, Aragón, Castilla La Mancha, Asturias, Cataluña sin Barcelona, Sevilla, La Rioja, Navarra, Comunidad Valenciana sin Valencia, País Vasco, Extremadura, Cantabria y Canarias. Sin embargo, sólo en Baleares y Murcia la probabilidad de declarar un siniestro es menor que en Andalucía sin Sevilla.

A continuación, la profesión del asegurado también permite jerarquizar las distintas categorías, en lo que al riesgo de siniestralidad se refiere. En este

sentido, se ha comprobado que los conductores con profesión 41[8] presentan una "odds" que es 1.885 veces mayor que los de la profesión 20[9], que se ha tomado como categoría de referencia (por ser la más frecuente). Aquellos conductores con código de profesión 22[10], 42[11] ó 64[12] también presentan un riesgo de siniestralidad mayor que los de la categoría de referencia, aunque las diferencias no sean tan significativas. Sin embargo, los conductores con profesión 53, 31, 57, 10, 21, 54, 55, 0, 12 ó 56[13] corren menos riesgo de tener un siniestro que los de la categoría de referencia (habiéndose ordenado la serie de menor a mayor significatividad), aún cuando las diferencias hayan resultado importantes en todos los casos. En particular, el riesgo de los asegurados con profesión 56 es .205 veces el riesgo de los asegurados de la categoría de referencia. El resto de las profesiones consideradas no han resultado significativas a la hora de explicar la ocurrencia de un siniestro.

Tabla 2: Resultados del modelo logit sobre la siniestralidad de una cartera de asegurados en el ramo de automóviles

Variable	Coeficiente	Wald	Ratio de Odds	Media Muestral
Hombres	-.205***	78.384	.815	.744
Estado Civil				
Soltero	Referencia			.202
Casado	.034	1.616	1.035	.779
Viudo	.152	2.657	1.164	.011
Divorciado	.134	1.289	1.143	.007

[8] FUNCIONARIOS Y ADMINITRATIVOS (desplazamiento profesional habitual urbano).

[9] INDUSTRIALES, COMERCIANTES, PROFESIONES LIBERALES (sin desplazamiento profesional habitual).

[10] INDUSTRIALES, COMERCIANTES, PROFESIONES LIBERALES (desplazamiento profesional habitual interurbano).

[11] FUNCIONARIOS Y ADMINISTRATIVOS (desplazamiento profesiónal habitual interurbano).

[12] CONDUCTOR DE CAMIÓN/VEH. INDUSTRIAL DE TERCEROS.

[13] Por orden: ESTUDIANTES, VIAJANTES Y REPRESENTANTES URBANOS, EMPLEADOS QUE CONDUCEN CON EXCLUSIVIDAD VEHÍCULOS DE LA SOCIEDAD, AGRICULTORES Y SUS EMPLEADOS, INDUSTRIALES, JUBILADOS, OBREROS, *códigos 0 y 12 "SIN CLASIFICAR"* , SIN PROFESIÓN.

Antigüedad del carnet	.001	.668	1.001	21.492
Tipo de permiso				
Coche	Referencia			.997
C	-.249	1.062	.779	.002
C1	-.998**	6.499	.369	.001
Motocicletas	.850	1.913	2.339	.000
Ciclomotor	.642	1.119	1.9	.000
D-D1-B2	.334	.300	1.397	.000
Edad del conductor	-.008***	60.124	.992	45.572
Antigüedad del vehículo	-.049***	534.287	.952	6.291
Valor en euros				
hasta 1050	Referencia			.359
1051-1350	.245***	63.225	1.278	.191
1351-1950	.283***	72.439	1.328	.280
más de 1950	.335***	64.040	1.398	.170
Potencia (en watios; 1CV≈736 watios)				
hasta 64002	Referencia			.257

Variable	Coeficiente	Wald	Ratio de Odds	Media Muestral
64003-85002	-.083***	7.973	.921	.246
85003-103002	-.015	.186	.985	.248
más de 103002	-.196***	26.705	.822	.249
Región				
Andalucía sin Sevilla	Referencia			.208
Sevilla	.511***	188.135	1.668	.084
Aragón	1.281***	499.724	3.599	.031
Asturias	.948***	299.361	2.579	.034
Baleares	-.191***	7.758	.826	.026
Canarias	1.017**	5.107	2.764	.000
Cantabria	.179**	6.125	1.196	.017

Castilla La Mancha	.796***	345.571	2.216	.060
Castilla y León	1.693***	1326.112	5.436	.057
Cataluña sin Barna	.863***	221.482	2.371	.029
Barcelona	2.008***	1743.791	7.450	.064
Extremadura	.312***	39.957	1.366	.040
Galicia	1.693***	1150.931	5.435	.049
Madrid	2.054***	2295.387	7.803	.094
Murcia	-.115**	4.480	.891	.037
Navarra	.885***	171.843	2.422	.020
País Vasco	.507***	114.079	1.661	.046
La Rioja	1.682***	177.363	5.374	.006
Com.Valenciana sin Val	.620***	132.487	1.859	.033
Valencia	1.005***	566.972	2.732	.062
Profesión				
20	Referencia			.433
0	-1.097***	997.077	.334	.093

Variable	Coeficiente	Wald	Ratio de Odds	Media Muestral
10	-.482***	69.003	.617	.026
12	-2.021***	1475.477	.133	.056
21	-.515***	90.273	.598	.030
22	.320***	7.259	1.377	.007
31	-.717***	32.774	.488	.005
40	.052	1.084	1.053	.038
41	.634***	384.728	1.885	.126
42	.460*	3.291	1.584	.002
53	-.595***	25.44	8.551	.006
54	-1.186***	116.454	.305	.009
55	-.788***	408.854	.455	.065

56	-1.587***	1873.128	.205	.100
57	-1.443***	63.914	.236	.003
63	.610	2.084	1.840	.001
64	.777*	3.216	2.175	.001
11-32-51	.135	.142	1.145	.001
Uso				
110	Referencia			.841
111	2.063***	79.403	7.866	.005
114	-.727	2.454	.483	.000
118	-.362***	20.146	.696	.014
119	2.187**	4.067	8.908	.000
131	2.450***	10.367	11.593	.000
141	3.143***	9.435	23.162	.001
150	2.340**	5.059	10.386	.000
160	.644***	49.160	1.903	.014
168	-.848***	158.812	.428	.024

Variable	Coeficiente	Wald	Ratio de Odds	Media Muestral
210	-.324***	29.648	.723	.027
211	1.041	.756	2.831	.000
212	1.986	1.878	7.285	.000
213	.579	.268	1.784	.000
217	-.355***	72.715	.701	.061
219	2.315**	4.866	10.122	.000
220	-.806**	4.461	.447	.001
228	-1.190	1.155	.304	.000
230	.556	.861	1.743	.000
231	2.421**	5.439	11.257	.001
232	1.732**	5.039	5.649	.000

238	-.401***	14.419	.670	.008
240	-.227	.534	.797	.001
Resto	3.706***	12.996	40.711	.001
Constante	.355***	47.210		
Chi cuadrado:	15592.702***			
R^2 de Nagelkerke:	.313			

Nota. La categoría Resto de la variable Uso es el resultado de solapar las siguientes categorías de dicha variable: 117, 133, 137, 190, 199, 218, 234, 235, 258 y 242.

* $.05 < p \leq .1$ ** $.01 < p \leq .05$ *** $p \leq .01$

La antigüedad del vehículo es la siguiente variable con más peso en el modelo. El riesgo de que un vehículo, con una antigüedad determinada, sufra un accidente es .952 veces el que tiene ese mismo vehículo un año antes.
El uso que el conductor hace del vehículo es la siguiente variable en importancia [14]. En este caso, la mayor parte de las categorías presentan mayor probabilidad de sufrir un accidente que la categoría 110[15] que se tomó de referencia (por resultar también la más frecuente). Las categorías con las diferencias más significativas (en orden descendente) son la 111[16], la 160[17], la categoría resto, la 131[18] y la 141[19]. Otros usos con mayor probabilidad de sufrir accidente que la categoría 110, pero con menor peso que los anteriores, son (ordenados de mayor a menor significatividad): 231, 150, 232, 219 y 119[20]. Por el contrario, los usos con código 220, 238, 118, 210,

[14] Algunas categorías de esta variable con frecuencias reducidas presentaban coeficientes estimados muy elevados, acompañados de desviaciones típicas también muy elevadas con niveles de significación bajos. Este comportamiento indicaba la presencia de multicolinealidad en el modelo (Greene, 1998; Hosmer & Lemeshow, 1989). Por lo tanto, se procedió a reunir las categorías afectadas (117, 133, 137, 190, 199, 218, 234, 235, 258 y 242) en otra codificada como *resto*. Después, desapareció este problema.
[15] TURISMO DE USO PARTICULAR.
[16] TURISMOS MATRICULADOS A NOMBRE DE EMPRESA.
[17] VEHICULO TODO TERRENO.
[18] TAXI SIN TAXIMETRO.
[19] VEHICULOS DE ALQUILER SIN CONDUCTOR.
[20] Por orden: FURGONETAS DE TRANSPORTE DE MERCANCIAS NO PELIGROSAS, TURISMOS DE AUTO-ESCUELA, AMBULANCIAS,

217 y 168[21] presentan menor riesgo de accidente que la categoría de referencia (estando ordenados de menor a mayor significatividad). En particular, el riesgo de declarar un siniestro por parte de un asegurado con uso 168 es .428 veces el que corre un asegurado de la categoría 110. La probabilidad de sufrir un siniestro por parte de aquellos asegurados que dan otros usos al vehículo no difiere significativamente de la de los asegurados con código de uso 110.

A continuación figura la variable valor del vehículo que, inicialmente, se trató con un carácter cuantitativo. Al correr la regresión, el parámetro que la pondera resultó significativo pero con una estimación de cero en sus tres primeras posiciones decimales. Una vez descartados problemas de multicolinealidad se decidió hacerla cualitativa. Para ello se procedió a establecer cuatro clases, en función de los cuartiles, y a volver a estimar el modelo con la variable cualitativa, tomando de referencia la categoría inferior (vehículos con un valor de hasta 1050 €). El resultado muestra que cuanto mayor es el valor del vehículo, mayor es la probabilidad de declarar un siniestro. En particular, los asegurados cuyos vehículos valen más de 1950 €, tienen un riesgo de accidente que es 1.398 veces mayor que el de aquellos que conducen vehículos de hasta 1050 €.

Por su parte, la condición de hombre reduce la probabilidad de accidente, siendo su riesgo 0.815 veces el que corre una mujer. Este resultado ya se produce cuando se cruzan las variables sexo y siniestro, por cuanto el porcentaje de hombres que declaran siniestro es 7 puntos inferior al de las mujeres (49.6% frente a 56.5%).

Asimismo, el riesgo de una persona con una edad concreta es .992 veces el de otra con un año menos.

El siguiente factor explicativo de la siniestralidad es la potencia del vehículo. Esta variable tuvo un comportamiento similar al de la variable valor, en el sentido de incluir una estimación de cero en el parámetro, a pesar de resultar significativa. En virtud de ello, se procedió de la misma forma, dividiendo la

---

FURGONETAS DE REPARTO URBANO Y AUTOVENTA, VEHICULOS DE SERVICIO DE URGENCIAS (POLICIA Y BOMBEROS)

[21] Por orden: FURGONETAS DE USO RURAL HASTA 500 KGS DE CARGA USO PROPIO - Uso 220, FURGONETAS USO PARTICULAR +5 HASTA 9 PLAZAS - Uso 238, TURISMOS USO PARTICULAR +5 HASTA 9 PLAZAS - Uso 118, FURGONETAS HASTA 3500 KGS - Uso 210, FURGONETAS HASTA 500 KGS DE CARGA USO PROPIO - Uso 217, TODO TERRENO +5 HASTA 9 PLAZAS - Uso 168.

variable en cuatro clases a partir de sus cuartiles. Los resultados muestran que los vehículos menos potentes presentan un riesgo de siniestralidad mayor. Ocurre lo contrario cuando se elimina la variable valor de la regresión, lo que demuestra que el comportamiento de la variable potencia viene determinado por el valor del vehículo (de hecho ambas variables presentan un elevado grado de correlación). La inclusión en la regresión de los dos factores permite aislar la verdadera influencia de la potencia del vehículo sobre el riesgo de sufrir un accidente.

Asimismo, al analizar la influencia de las diferentes modalidades de permisos de conducción existentes, sólo ha resultado significativo que aquellos conductores con permiso de circulación tipo C1 corren menos riesgo de sufrir un accidente que los que poseen un permiso de circulación de coche.

No se han apreciado diferentes niveles de riesgo según el estado civil del conductor. Tampoco la antigüedad del carnet ha resultado significativa. Esta circunstancia se explica por la elevada correlación que mantiene esta variable con la edad del conductor (de hecho, aquella aparece con un beta de -.006 y un Wald de 24.657 cuando esta se retira del modelo) y la menor influencia que tiene sobre la siniestralidad, lo que hace que sea la edad del conductor la variable que resulte significativa.

De las 105 variables de entrada utilizadas, 58 han resultado significativas (46 con un p-valor inferior al 1%, 10 con un p-valor entre el 1 y el 5% y sólo 2 con un p-valor superior al 5%).

Resultados obtenidos por el Algoritmo Genético

Para escoger los factores que retenían mayor información de cara a la siniestralidad, se programó un AG al efecto, utilizando el cruce en un punto y la mutación de un solo bit. El objetivo era escoger los 30[22] factores con mayor predictivo. La función de capacidad era la capacidad predictiva estimada con una SVM. Observamos los resultados en la siguiente tabla:

[22] Es un número obtenido por *búsqueda generacional*, y está relacionado con el nº de datos y el de variables.

Tabla 3: Los 30 factores seleccionados por el AG

1. Antigüedad carnet	2. Edad conductor
3. Antigüedad vehículo	4. Madrid
5. Barcelona	6. Valor
7. Potencia	8. Castilla y León
9. Galicia	10. Profesión 12
11. Profesión 41	12. Casado
13. Hombres	14. Uso110
15. Andalucía sin Sevilla	16. Profesión56
17. Murcia	18. Aragón
19. Soltero	20. Sevilla
21. Profesión 20	22. Baleares
23. Profesión 0	24. Valencia
25. Cataluña sin Barcelona	26. Castilla La Mancha
27. Comunidad Valenciana sin Valencia	28. Uso 168
29. Asturias	30. Profesión 55

Comparación de los resultados obtenidos por ambas técnicas

En lo que respecta a la identificación de las variables que explican la siniestralidad, el modelo logit ofrece una información más rica que el AG. Aquel, no se limita a informar sobre qué factores inciden en la siniestralidad y con qué peso, sino que además permite conocer el sentido de dicha influencia.

Al comparar los resultados obtenidos se observan diferencias notables entre ambos modelos. Así, aunque la mayor parte de los factores del AG se encuentran seleccionados por el logit, el que tiene más importancia en aquella –que no es otro que la antigüedad en el carnet– no resulta siquiera significativo en este. No parece, a este respecto, que el AG esté recogiendo la elevada correlación existente entre este factor y la edad del conductor que también resulta seleccionada. Melgar y Guerrero (2005) también seleccionaron simultáneamente edad y antigüedad como variables significativas después de aplicar un modelo econométrico tipo *count data*; eso sí, la última variable con carácter dicotómico (menos de 2 años; 2 años o más).

El estado civil tampoco parece tener suficiente peso en el modelo logit, mientras que soltero y casado sí lo tienen en la selección dada por el AG. Al

considerar en el modelo logit los factores más significativos (con mayor Wald) y compararlos con los respectivos del AG (véase tabla 3), se observa que el lugar geográfico, la profesión y la antigüedad del vehículo son las variables con más peso en dicho modelo, mientras que en el AG la antigüedad del carnet, la edad del conductor y la antigüedad del vehículo tienen más importancia que el lugar geográfico o la profesión. También se observa que hay mayor diversidad entre los 30 factores seleccionados por el AG, que entre aquellos que tienen más peso en el modelo logit.

4.3 Clasificación tras la selección

Comparemos por último la tasa de clasificación que se alcanzó tras ejecutar de nuevo la SVM sólo con los anteriores factores de riesgo, frente a la que se obtuvo en el punto 4.1 con todos los factores.

Tabla 4

Clasificación	30 variables seleccionadas por el AG	Todas las variables
SVM	77.66%	77.72 %

Como se puede apreciar, el porcentaje de casos bien clasificados es prácticamente el mismo utilizando los 30 factores de riesgo que empleando todos los recogidos por la aseguradora. Éste resultado es francamente interesante para la Compañía, de cara al ahorro en tiempo y recursos a la hora de recoger esos datos. El haber eliminado información redundante para el sistema hará, por otra parte, que mejore la estabilidad del clasificador, y que el coste computacional del proceso sea menor.

Conclusiones

En el presente artículo se han introducido técnicas novedosas, tomadas del Aprendizaje Máquina, para resolver un problema de tarificación a priori. Concretamente, se trata de clasificar un grupo de asegurados descritos por sus factores de riesgo en dos clases, atendiendo a si presentan o no siniestro en el periodo de un año. Este problema engloba una segunda cuestión, la de la selección de los factores de riesgo que mayor información recogen de cara a la siniestralidad.

La aplicación de una SVM y un AG, y un modelo logit al análisis de la siniestralidad ha ofrecido cierta coincidencia respecto a los factores seleccionados, aunque no en cuanto a su ordenación. La primera técnica ha resultado más eficaz en lo que al porcentaje de acierto en el pronóstico se refiere, como también ocurrió con el Análisis Discriminante. Sin embargo, el logit tiene la ventaja de poder estimar el aumento o la disminución del riesgo de siniestralidad, ante un cambio en uno de los factores. En este sentido, podría pensarse en utilizar un AG para seleccionar los factores y un logit para conocer el sentido de la influencia de cada uno de ellos, siendo conscientes de que no será posible tal conocimiento en aquellos factores que no resulten significativos en el logit.

Concluimos con una nueva clasificación utilizando sólo los 30 factores seleccionados por el AG, en la que observamos cómo se alcanza prácticamente la misma tasa que si clasificamos atendiendo a todos los factores de riesgo recogidos por la aseguradora.

Durante el curso de nuestras investigaciones, hemos realizado otros experimentos, seleccionando factores con árboles de clasificación, con conclusiones similares.

Más interesantes han resultado las prácticas que hemos realizado con una nueva base de datos, correspondiente al año 2005. En este caso, los factores de riesgo que había recogido la aseguradora eran muy diferentes a los que aparecían en la base de 2005, e incluían una variable, el nivel de *Bonus Malus*, que nos llevó a la ejecución de experimentos levemente diferentes. Se trataba de ver la influencia que tenía a la hora de seleccionar factores y de clasificar. Por ello, realizamos una selección con un *Random Forest* incluyendo y sin incluir esta variable en la entrada, y una clasificación de los asegurados con el mismo criterio. En la selección, se vio claramente que la influencia del nivel de Bonus Malus era considerable, resultando escogidos muchos de estos niveles. La clasificación, por encima del 70 % en ambos casos, resultaba mejor si se añadía el nivel de Bonus Malus que en caso contrario.

Bibliografía

Aldrich, J. & Nelson, F.D.. Linear Probability, Logit and Probit Models. Sage University Papers: Quantitative Applications in the Social Sciences. Berverly Hills: Sage Publications, 1986.

Bousoño Calzón, Heras Martínez, Tolmos Rodríguez-Piñero. Factores de Riesgo y Cálculo de primas mediante Técnicas de Aprendizaje. FUNDACIÓN MAPFRE, Madrid, Junio 2008.

Burges, C. J.: "A tutorial on Support Vector Machines for pattern recognition", Knowledge Discovery and Data Mining, 2(2) (1998):121-167.

Duda, R.O., Hart, P.E. Stork, D.G. (2001) *Pattern Classification*. John Wiley & Sons.

Chih-Chung Chang and Chih-Jen Lin, *LIBSVM : a library for support vector machines*. Software disponible en www.csie.ntu.edu.tw/~cjlin/libsvm.

Greene, W.H.. Análisis Econométrico. Prentice Hall, 1998.

Heras Martínez, Bousoño Calzón, Tolmos Rodríguez-Piñero, Santiago Mozos. Selección de Factores de Riesgo y Predicción de Siniestros en el Seguro del Automóvil Mediante Métodos de Aprendizaje Máquina. Actas del Congreso RIESGO 2007.

Hernández-March, J. La Emancipación Juvenil: Un Análisis Estadístico Aplicado a la Comunidad de Madrid. [Recurso electrónico] Tesis Doctoral. [Madrid]: Universidad Complutense de Madrid, Servicio de Publicaciones, 2003.

Hernández-March, J., Tolmos, P. Un análisis comparativo de una SVM y un modelo Logit en un problema de clasificación de asegurados. Actas del Congreso RIESGO 2009.

Hosmer, D.W. & Lemeshow, S.. Applied Logistic Regression. New York: John Wiley (1989).

Liao, T.F.. Interpreting Probability Models: Logit, Probit and Other Generalized Linear Models. Sage University Papers: Quantitative Applications in the Social Sciences. Thousand Oaks: Sage Publications, 1994.

Melgar-Hiraldo M.C., Guerrero-Casas, F.M. Los siniestros en el seguro del automóvil: un análisis econométrico aplicado. Estudios de Economía Aplicada. Abril 2005 vol.23, número 001.

Salas-Velasco, M.. Graduates on the labor market: Formal and informal post-school training investments. Higher Education, 54(2), 227-246, 2007.

Salcedo-Sanz S.,De Prado-Cumplido M., Pérez-Cruz F., Bousoño-Calzón C. *Feature Selection via Genetic Optimization.* ICANN 2002: 547-552.

Salcedo-Sanz, S., Fernández-Villacañas J. L., Segovia-Vargas, M. J. and Bousoño-Calzón, C. 2005. Genetic programming for the prediction of insolvency in non-life insurance companies, Computers & OR 32: 749-765.

Segovia-Vargas MJ, Salcedo-Sanz S, Bousoño-Calzón C. Prediction of insolvency in non-life insurance companies using Support Vector Machines and genetic algorithms. In: Proceedings of X SIGEF Congress in Emergent Solutions for the Information and Knowledge Economy, León, Spain, 2003.

Shapiro, A. The merging of Neural Networks, fuzzi logic and genetic algorithms. 2002, Insurance: Mathematics and Economics 31 (2002) 115-131.

Tolmos Rodríguez-Piñero, P. Selección de Factores de Riesgo y Predicción de Siniestros en el Seguro del Automóvil Mediante Métodos de Aprendizaje Máquina. Tesis Doctoral. Universidad Rey Juan Carlos. Madrid 2007.

Vapnik V., Chervonenkis A. 1974. *Theory of Pattern Recognition* (*in Russian*) Nauka, Moscú.

Vapnik, V., Cortes, C. *Support-Vector Networks.* 1995. Machine Learning, 20, 273-297.

Weston, H., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V. (2000): *Feature Selection for SVMs*, Advances in NIPS 12, MIT Press, 526-532.